Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling

Year: 2024 | Citations: 1035 | Authors: Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao

---

Abstract

We introduce InternVL 2.5, an advanced multimodal large language model (MLLM) series that builds upon InternVL 2.0, maintaining its core model architecture while introducing significant enhancements in training and testing strategies as well as data quality. In this work, we delve into the relationship between model scaling and performance, systematically exploring the performance trends in vision encoders, language models, dataset sizes, and test-time configurations. Through extensive evaluations on a wide range of benchmarks, including multi-discipline reasoning, document understanding, multi-image / video understanding, real-world comprehension, multimodal hallucination detection, visual grounding, multilingual capabilities, and pure language processing, InternVL 2.5 exhibits competitive performance, rivaling leading commercial models such as GPT-4o and Claude-3.5-Sonnet. Notably, our model is the first open-source MLLMs to surpass 70% on the MMMU benchmark, achieving a 3.7-point improvement through Chain-of-Thought (CoT) reasoning and showcasing strong potential for test-time scaling. We hope this model contributes to the open-source community by setting new standards for developing and applying multimodal AI systems. HuggingFace demo see https://huggingface.co/spaces/OpenGVLab/InternVL